Филонов Алексей Петрович, аспирант
Научный руководитель: Доросинский Леонид Григорьевич, проф., д-р техн. наук

ПОЛУЧЕНИЕ ПОМЕХОУСТОЙЧИВЫХ ПРИЗНАКОВ ОПИСАНИЯ РЕЧЕВОГО СИГНАЛА В СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

Речевой сигнал содержит в себе много избыточной информации, поэтому в системах распознавания речи часто используют методы снижения размерности исходного сигнала в некоторое компактное множество параметров. Поэтому от того насколько надежные параметры получены на этапе предобработки сигнала, зависит насколько удачным будет процесс обучения системы распознавания речи, а следовательно, в дальнейшем и сам процесс распознавания.

Считаем, что акустический сигнал это дискретная последовательность отсчетов. Исходный акустический сигнал подвергается блочной обработке, блоками 10 – 20 мс. Именно на участке такой длительности речь считается квазистационарной и может быть привязана к конкретной реализации единицы речи (фонеме). Таким образом, мы имеем $\{s_1..s_k\}$ отсчетов в блоке. Выделенный блок (последовательность отсчетов) предварительно подвергается обработке, оконной функцией (1) (обычно используется окно Хэмминга, но, впрочем, может быть и любое другое). Целью данного этапа обработки является снижение граничных эффектов, возникающих в результате сегментации.

$$y_i = s_i w_i$$

$$w_i = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi i}{K-1}\right) & 0 \le i < K \\ 0 \quad i < 0, i \ge K \end{cases}, \tag{1}$$

Далее производится перевод сигнала в спектральную область с использованием дискретного преобразования Фурье. При этом в дальнейшем используется лишь амплитудно-частотная характеристика, поэтому фазо-частотную характеристику можно отбросить.

В конкретный момент времени при наличии в сигнале аддитивной помехи

$$x_i = y_i - n_i, \tag{2}$$

где y(i) – зашумленный сигнал, x(i) – исходный (чистый) сигнал, n(i) – аддитивная помеха.

В связи с тем, что преобразование Фурье обладает линейным свойством (принцип суперпозиции), то в частотной области мы получаем выражение (3)

$$|X| = |Y| - |N|, \tag{3}$$

где |Y|, |X|, |N| – амплитудные спектры зашумленного, исходного сигналов и аддитивной помехи.

Участки акустического сигнала могут быть разделены на два вида: вокализованные (речь присутствует) и невокализованные. На невокализованных участках возможно получить усредненную оценку амплитудного спектра шумовой составляющей.

$$|\overline{N}| = \frac{1}{T}\sum_{t=1}^{T}|N(t)|, \quad (4)$$

где Т – количество участков, используемых для получения средней оценки амплитудного спектра шума.

Тогда при условии, что помеха остается постоянной либо медленно меняющейся до следующего невокализованного участка, на котором будет обновлена усредненная оценка, можно записать следующее выражение:

$$|\hat{X}| = |Y| - |\overline{N}| = |Y| - \frac{1}{T}\sum_{t=1}^{T}|N(t)|. \quad (5)$$

Помимо этого подхода можно использовать винеровскую фильтрацию.

Фильтр с конечной импульсной характеристикой во временной области описывается выражением:

$$x(l) = h(l) * y(l), \quad (6)$$

где l – порядок фильтра.

В частотной области свертка преобразуется в произведение (7).

$$X = H \cdot Y. \quad (7)$$

Следовательно, мы можем на невокализованных участках оценить коэффициенты фильтра с использованием усредненной оценки спектра шумовой составляющей (8).

$$\hat{H} = \frac{X}{Y} = \frac{Y - \overline{N}}{Y} = 1 - \frac{\overline{N}}{Y}. \quad (8)$$

Далее мы производим фильтрацию зашумленного сигнала в частотной области.

$$\hat{X} = \hat{H} \cdot Y. \quad (9)$$

В качестве признаков описания обычно используется кепстральное описание, получаемое обратным преобразованием Фурье логарифма амплитудного спектра сигнала. Для снижения размерности признаков предварительно применяется нелинейное сжатие в мел-или барк-шкалу. Окончательно мы получаем набор мел (или барк) кепстральных параметров, которые могут быть использованы для работы системы распознавания речи (обучения и непосредственно распознавания).